DOCUMENT RESUME

ED 106 337 95 TM 004 452

AUTHOR Everett, Bruce E.
TITLE Effective Data Management and Quality Control Techniques for Large-Scale Longitudinal Research.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE [Apr 75]
CONTRACT OEC-0-70-4789
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 1, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Computer Programs; *Data Collection; *Educational Innovation; Electronic Data Processing; *Longitudinal Studies; *Management Systems; *Methods; Public Schools; Recordkeeping; Research Problems

ABSTRACT

This paper summarizes the data management techniques used in a large-scale longitudinal study of intensive, innovative programs. Emphasis is placed on the scope of data collection, the methods used in the study to keep track of students over several years, the techniques used to collect data about students in public schools, the quality control on the data, and the data-file management system used to reference and cross-reference multilevel variables for subsequent analyses. (Author)

10,01

EFFECTIVE DATA MANAGEMENT AND QUALITY CONTROL TECHNIQUES FOR LARGE-SCALE LONGITUDINAL RESEARCH*

Bruce E. Everett

American Institutes for Research
Palo Alto, California

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education and Welfare (Contract Number OEC-0-70-4789). Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

*A presentation given at the 1975 annual meeting of the American Educational Research Association, Washington, D. C.

Effective Data Management and Quality Control Techniques for Large-Scale Longitudinal Research

Bruce E. Everett

American Institutes for Research

Introduction

The very nature of longitudinal studies in education creates a number of methodological problems in the tracing of students through their educational careers. Moreover, any large data collection effort in the public schools, whether it is longitudinal or not, will encounter many obstacles which were not envisioned during the planning stages of such a study. The purpose of this paper is to review a data management system which was designed specifically to deal with these concerns. It was used in AIR's LONGSTEP, a Longitudinal Study of Educational Practices, which involved nearly 30,000 students and more than 1,500 teachers in 80 schools over a three-year period.

Identification of Students

The first requirement in a longitudinal study is that it apply a method of reliably and conveniently identifying each and every student when data are collected, so that information from several time points or several data collection instruments can be successfully merged together for any given student. From the data manager's standpoint, the ideal common denominator for all forms is a unique student identification number, preferably with several digits allocated to specify the school or at least the district in which the student resides. If every student were correctly identified through the existence of his or her ID number on all instruments, then the process of building a master data file containing all student level data collected during the course of a study would be comparatively simple. However, there are many reasons why this does not always happen in the field:

some students never receive their assigned number; others leave in mid-test and make up the rest on another form; some get the wrong ID number but use it anyway; and some will inadvertently or deliberately alter the digits of the ID number assigned to them.

One of our objectives during LONGSTEP was to minimize the confusion caused by student mobility. At the start of each school year, an elaborate chart was prepared which showed the participating grades within each school and where the students within them had been the previous year. This information was successfully used to anticipate where the majority of our participating students would be even as they moved from elementary school to junior high.

Requiring that every student fill out complete identifying information on every data collection form during LONGSTEP would have placed a needless burden on those students and would have been very difficult to implement at the primary grade levels. Moreover, many inconsistencies (use of nicknames, ignorance of birthdate, etc.) would have resulted from such a procedure. As a means of identifying student level data with a minimum of confusion and expense, computer generated, stick-on labels were used which contained all pertinent ID information for each student.

Instructions relating to the placement of labels on data collection instruments specified that new students, or old students who did not receive their proper labels, were to fill out the identification information on the forms themselves. ID numbers were then assigned to such students in-house before raw data forms were optically scanned or keypunched. Records of who was assigned what number were maintained in order to facilitate our subsequent data processing, then stored for eventual destruction. Before each administration, the labels were sorted into packets by AIR

staff according to testing groups specified by each site. The labels were placed on the forms in the field.

Preparation and Scoring of Raw Data Forms

The preparation of data forms prior to their conversion to computer-readable records involves several basic operations. First, problems with the forms themselves (incorrect ID's, wrinkled paper, unwanted pencil marks, poor erasures, etc.) should be corrected. Second, the forms should be arranged in systematic order (lowest to highest ID within grade within school, for example); and this order, together with the number of forms, should be marked on the outside of the boxes the forms are packed in. This serves two purposes: not only can one compare the number of forms sent out with the number returned from the field, but one is also able to specify a precise order in which the data are subsequently processed.

Even when the data come back from keypunch or optical scoring, there are likely to be minor changes which are desirable to make. Late arriving forms, duplicate records (a not-uncommon problem when keypunching), and the like, can be most efficiently handled through the use of computer terminals and text-editing routines. One can create, delete or modify individual records in an online disk data set before consigning it to cheaper but less accessible tape storage.

Once all of the basic raw data have been converted into a series of computer-accessible data sets, the process of comparing them and merging them together can begin. If one has achieved a consistent set of ID numbers for all individuals at all times, this will not be very difficult; on the other hand, problems of misidentification will impact on each and every data set being used. For reasons of cost-effectiveness, then, it is highly advisable to clean up all problems of misidentification before actually merging data together to create a final data base.

The Data Management System of LONGSTEP

During the design phase of LONGSTEP it became evident that a comprehensive and flexible data management system would be the key to effective subsequent analysis of the data. To store, retrieve, and cross-reference extensive information on nearly 30,000 students, approximately 1,500 teachers and the hundreds of educational treatments occurring in the participating districts necessitated a very sophisticated data handling system. An essential requirement of the data handling system was that each student's performance data, background, attitudes, and exposure to educational treatments had to be tied to that student for each year of LONGSTEP. In addition, teacher and educational treatment data had to be directly linked with each individual student having that teacher and treatment.

In order to maintain the necessary information about the students, teachers and educational treatments, several major data files had to be developed as part of the data management system. A student master file had to be generated which would contain all performance data, background, and attitudes collected for each student during the course of LONGSTEP. A teacher/treatment master file containing teacher characteristics and educational treatments also had to be generated and maintained (see Figure 1). The method of cross-referencing between these two files will be discussed later.

An integral part of LONGSTEP's data management system was a computer program which could take each input data record and systematically inspect, edit and insert it into the appropriate master file. This computer program allowed for the creation and maintenance of a given data file composed of separate records representing unique entities.

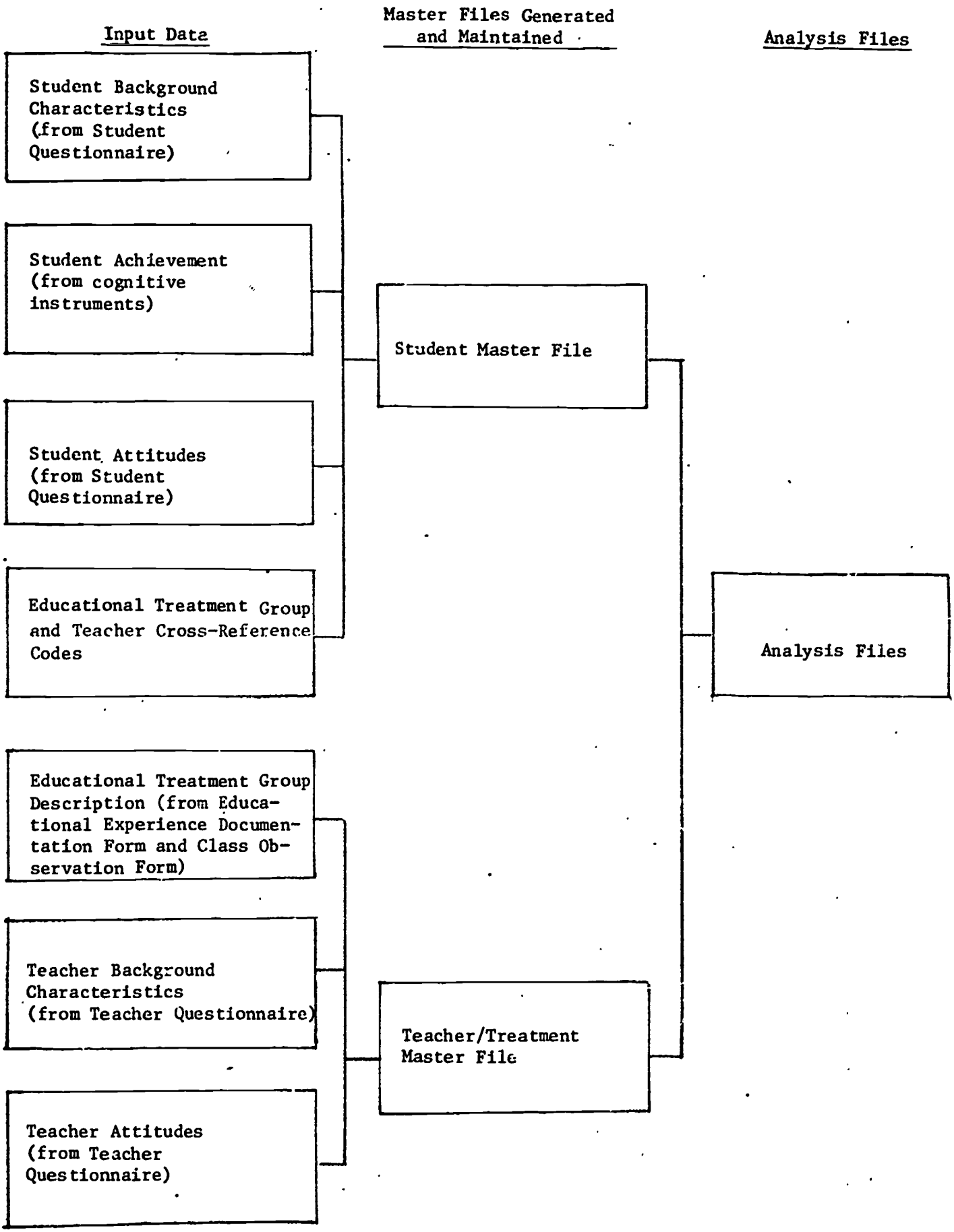

Figure 1. General data management system structure of LONGSTEP.

All data stored with respect to each student, teacher or treatment were recorded as values for specific variables, each of which was user-named at the time the data/variables were added to the file.

The flexibility of the file maintenance program was evidenced by the fact that files were created and updated from input data records possessing a wide variety of formats. In fact, the only practical constraints in creating or updating files was that one had to specify: (a) the names of each variable to be added; (b) the column location in the input records where the score for that variable was located; (c) an identifier on each input data record, if input records added to the file on a single run differed with respect to the type of data they contained; and (d) the name, location and configuration of input and output files.

The file maintenance program required that each input record so processed had to contain the complete entity identification (ID) code. If the ID of an input record matched the ID of an entity already in the file, the input data were added to the file as part of the entity already there. This occurred, for example, when new test and questionnaire data were added to the student master file in subsequent years. On the other hand, if the input ID did not match any of those in the file, a new entity was created in the file. Thus, the file maintenance program had the potential of creating new entities or records in a previously created file of data, adding to records already in that same file, and accomplishing both of these functions in a single run.

Since the program printed out all new entities added to a master file, it was a routine task to check for ID errors by comparing the ID's of new entries with those previously entered in the master file. When the first master file was created, all entries were new and all were therefore printed

out. Succeeding master file updates, then, always led to printouts which could be compared with a previously obtained printout. Also, all commands, comments and successful operations were permanently retained with the master file and were listed upon each file update.

The file maintenance program was also capable of editing input data. A range of editing limits for each variable being entered in the file was specified. If an item of input data failed the edit limits, it was treated as an error and maintained as a blank in the file. Missing data were maintained as blanks in the input data.

A separate data extraction program allowed the outputing of certain data from selected records in either or both of our master files. Records that met the selection criteria had specified data items (variables) formatted onto an output record. Output records could then be routed to the printer, to the card punch, or to any peripheral data set specified by the user. It is important to note at this point that the extraction program selected a given record only once, even if the record met a number of disjunctive (either-or) criteria.

The above paragraphs summarize the salient characteristics of the file maintenance and data extraction programs as they functioned in LONGSTEP's data management system. To these characteristics might also be added the rather significant quality, "simplicity of usage." Although the system could not be used effectively by an individual without knowledge of data processing techniques and computer job control language, the degree of sophistication required of a user was quite minimal. This is a rather important characteristic since it meant that research personnel themselves were able to interact directly with the data base.

Master File Cross-referencing

The design of LONGSTEP specifically stated that all teacher and educational treatment variables associated with a given student would involve characteristics of only those teachers and those treatments to which that particular student had been exposed. Achieving this degree of precision was complicated in that: (1) many students were exposed to more than one teacher or educational treatment, and (2) a group of students who had the same teacher(s) and educational treatment for one subject might not have had the same teacher(s) and educational treatment for other subject matter areas. Such variation in the manner in which individual students were exposed to teacher and educational treatment characteristics meant that it would be impossible to associate treatment and teacher data with students on the basis of grade and school membership information alone. Information specifically identifying each student's teacher(s) for each subject was gathered to provide this link.

A two-file system has developed to provide these cross-referencing capabilities. All student data were maintained in one file and all teacher and treatment data in a second file. The smaller of these two, the teacher/treatment file, contained teacher questionnaire data for teacher entities and educational treatment data for treatment entities. Data in the student file included all achievement test scores and all questionnaire data obtained from each student over the course of his participation in the study. This file also contained a series of cross-reference codes identifying each student's teacher(s) and educational treatment for each subject for each semester of participation in the study.

These teacher and treatment codes became the items in the student file that allowed the identification of the teacher and treatment entities

in the teacher/treatment file appropriate for each student. The crucial aspect of this cross-referencing process was that the appropriate information in the teacher/treatment file was contained in a record with an ID corresponding to the cross-reference code in the student file. Accurate teacher cross-referencing was assured by using the same teacher number to identify teacher questionnaires and to codify each student's teacher(s). Accurate treatment cross-referencing was assured by creating treatment entities in the teacher/treatment file which corresponded to the different treatment cross-reference ID's that were found in the student file.

This particular procedure allowed the development of treatment identification codes from teacher codes but did not require that they involve the same subsets of students. This methodology, then, achieves one of the data management goals of the study by permitting treatment and teacher data to be independently associated with students. The fact that teacher and treatment characteristics can be associated with each student so as to reflect the student's own educational environment is novel to large-scale educational studies. It is expected that this rather precise assessment of the school environment will permit student-level analyses which will yield significant new insights into longitudinal educational effects.